Deliverable

D9.4 Integrated Analytical Workflows

Integration and deployment of COMPARE integrated analytical workflows into a COMPARE computational workflow environment for autonomous sequence data analysis and submission

Version: 1.0

Due: 48

Completed: 48

Authors: Peter Harrison[1], Blaise Alako[1], Nima Pakseresht[1], Nadim Rahman[1], Martin Thomsen[2], Judit Szarvas[2], Ole Lund[2], Matt Cotten[3], David Nieuwenhuijse[3], Marion Koopmans[3], Basil Xavier[4], Surbhi Malhotra-Kumar[4], Dirk Höper[5], Martin Beer[5], Dennis Smitz[6] and Guy Cochrane[1]

[1]EMBL-EBI, UK.

[2]National Food institute Technical University of Denmark, Denmark.

[3]Erasmus Medical Center, The Netherlands.

[4]Laboratory of Medical Microbiology, Vaccine & Infectious Disease Institute, University of Antwerp, Belgium.

[5]Friedrich-Loeffler-Institut, Germany.

[6]National Institute for Public Health and the Environment (RIVM), The Netherlands

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 643476.

Contents

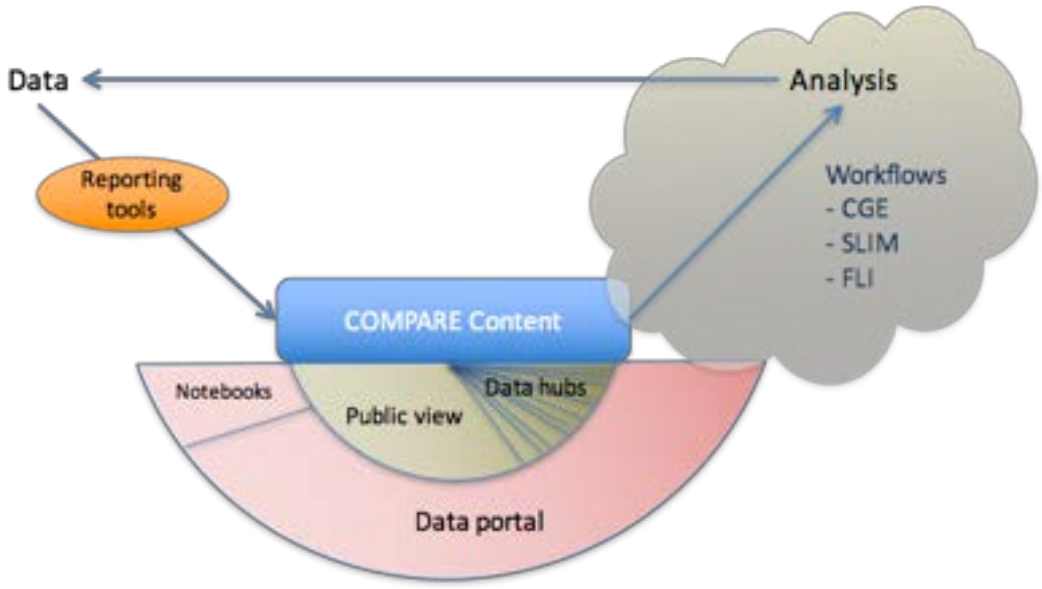
Introduction to report

In this report, we provide an overview of the COMPARE integrated analytical workflows that are being developed and subsequently deployed within the computational workflow engine developed by work package 9. The generic workflow engine itself is not the subject of this report, and its operation was documented within Deliverable 9.3. This report covers the actual computational analysis workflows that are already, or planned in the near future, to be deployed into the workflow engine. Once fully integrated, the generic workflow engine automates the operation of the workflows. This report provides an overview of the different analytical workflows on offer, the flow of data within them, the work undertaken to integrate them into the generic workflow engine, and a report on the output from the three analytical workflows that are already integrated into the workflow engine. We also report on our progress to integrate a further three computational analysis workflows into the generic workflow engine and further plans for development of the workflows within the COMPARE computational workflow environment.

Integrated Analytical Workflows

The generic workflow engine, that we have termed SELECTA, has been developed to ease the bottleneck of COMPARE-specific analysis and subsequent submission of results to the European Nucleotide Archive (ENA). The analysis component is a key part of the COMPARE data platform (Figure I), with the SELECTA workflow engine autonomously triggering computational analysis through one of several cloud-hosted workflows. SELECTA then manages the processing and submission of the output of these integrated analytical workflows, which mitigates the burden that a submitter faces when handling bulk submissions to the public archives and COMPARE data portal.

Figure 1: The COMPARE data platform showing major components.



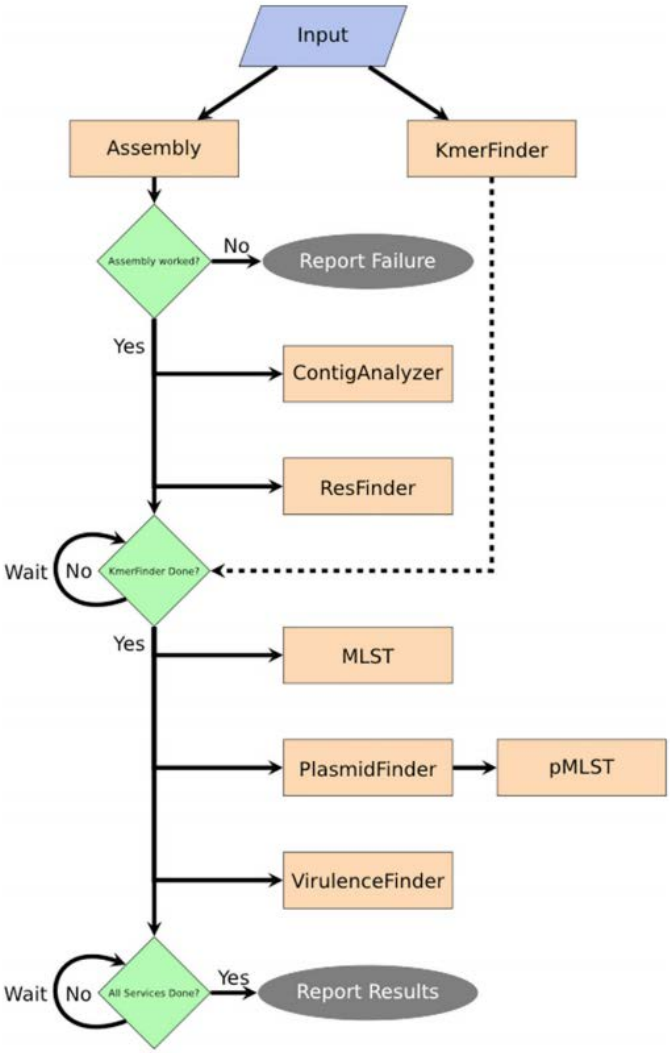
SELECTA consolidates analysis pipelines within the COMPARE consortium into a single computing and COMPARE-data access environment. The SELECTA system processes and analyses data via logical stages, namely data provision, core execution and analysis

submission. The private datahub reads are acquired at various grouping levels depending on set criteria. The following pipelines are fully integrated and functional in the SELECTA system. The Erasmus Medical Center Virus analysis pipeline (EMC SLIM); The Center for Genomic Epidemiology bacterial genome analysis pipeline of the Technical University of Denmark (DTU CGE); and The University of Antwerp bacterial genome analysis pipeline (UAntwerp BacPipe). We are currently actively working at integrating RIEMS, the Reliable Information Extraction of Metagenomic Sequence pipeline for comprehensive taxonomic classification of metagenomic sequence reads from the Friedrich Loeffler Institut, and Crypto Parasite pipeline, a pipeline for analysis of pathogen strains structural variations. We will now briefly summarize the integration of these five workflows into the SELECTA workflow engine.

DTU CGE

The Center for Genomic Epidemiology bacterial genome analysis pipeline of the Technical University of Denmark (DTU CGE) is a bacterial analysis platform for clinical diagnostics and surveillance that integrates several published state of the art web-based tools. The pipeline automatically identifies bacterial species from multiple bacterial isolates and if necessary assembles the genomes to identify multilocus sequence type, virulence gene and antimicrobial resistance gene. In the version 1.1 update, CgMLST and Salmonella Type finder capabilities were added in preparation for a trial by the European Centre for Disease Prevention and Control (ECDC). The pipeline provides a summary report of the analysis that enables a rapid overview, and results have performed well against benchmark datasets generated previously. The pipeline correctly predicts the species and most expected genes, automatically exploiting whole-genome sequencing to assist clinical diagnostics and pathogen surveillance. Figure 2 depicts the full workflow from raw Illumina read input to results summarised in a tab separated file. It is available from within the SELECTA framework but also the latest development builds directly from the repository (https://bitbucket.org/genomicepidemiology/cge-tools-docker/src/master/). SELECTA only updates when there is a major stable release, and will upgrade to the latest stable build (Version 1.2) when it is ready for release by DTU later in 2019. This will trigger a reanalysis of all existing data and a suppression of the analysis results produced by Version 1.1. The workflow is described in Thomsen et al. (2016) A Bacterial Analysis Platform: An Integrated System for Analysing Bacterial Whole Genome Sequencing Data for Clinical Diagnostics and Surveillance. PLoS ONE 11.

Figure 2. DTU CGE workflow for bacterial clinical diagnostics and surveillance.

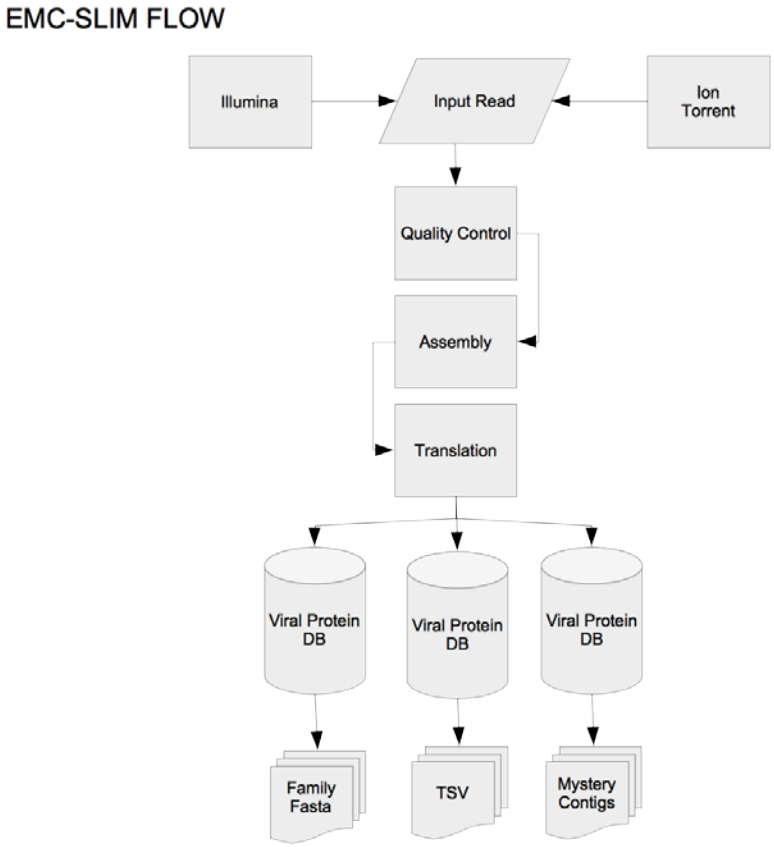


EMC SLIM

SLIM is a bioinformatics tools wrapper for the de novo assembly of Illumina HiSeq paired-end or Ion Torrent single end reads to contigs and contig classifications. SLIM initially performs a generic quality control of short read sequences by removing common adapters and low quality reads. A de novo assembly using SPAdes is performed on the set of reads that passed the quality control. SLIM by default uses generic parameters that provide a reasonable de novo assembly performance across a broad range of sequence data types and is therefore a useful starting point for further analysis. Furthermore, SLIM classifies the contigs based on a translation of all six reading frames of the contig and screens for homologies to viral proteins. A single tab-separated value output table is generated listing each contig and showing protein homology above 30% to sequences in the virus family-based database. All other intermediates files are provided as a compressed directory. While the results table and compressed intermediate files are available for inspection or download within the Pathogen Portal, we have also provided a dedicated browser for the results of EMC SLIM analysis, the ViromeBrowser. The ViromeBrowser is available as an R package and Shiny application from https://CRAN.R-project.org/package=viromeBrowser,

with user documentation at
https://cran.r-project.org/web/packages/viromeBrowser/vignettes/viromeBrowser.html.

Figure 3. EMC SLIM workflow for quality control, de novo assembly and classification.



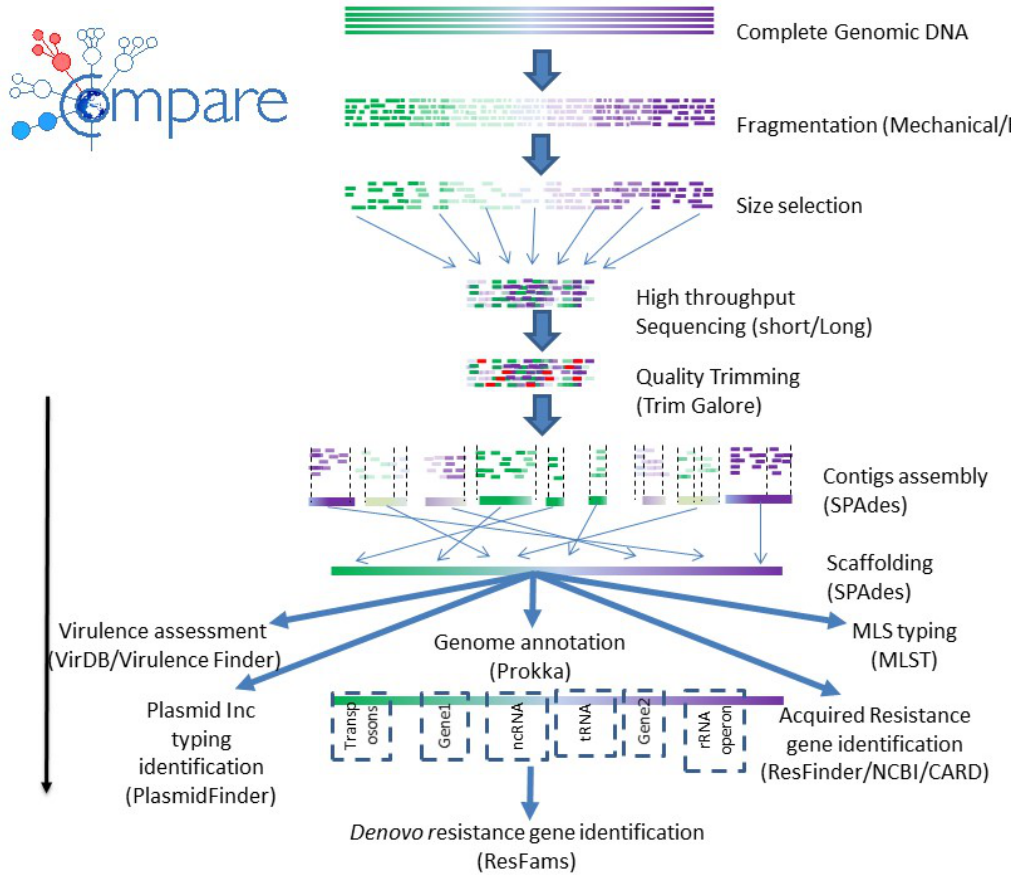
BACPIPE

BacPipe, a bioinformatic pipeline for bacteriology clinical diagnostic and outbreak detection was developed to mitigate the considerable bioinformatics expertise requirement for bacterial genome analysis. It is an ensemble of open-access bioinformatics tools combined into a complete workflow for the analysis of bacterial whole-genome sequencing. The reads are quality checked before genome assembly, annotation and identification of resistance and virulence genes. Additionally, BacPipe can simultaneously analyse several strains of bacteria to understand their evolutionary relationship and derive a bacterial transmission route. BacPipe was validated using published Methicillin-Resistant Staphylococcus Aureus (MRSA) outbreak data. The BacPipe tools consume Illumina, Pacific Pacbio, Nanopore, Sanger and Ion Torrent single or paired end reads and the output analysis results are consolidated into a single tabular sheet for rapid overview and interpretation by the microbiologist/clinician. BacPipe has key clinical applications in its facilitation of the analysis and interpretation of pathogen sequence datasets. Easing application to routine patient care within a hospital setting and the monitoring of infection control in public health.

BacPipe is fully integrated into the SELECTA environment and the software can be obtained from GitHub (https://github.com/basilbritto/bacpipev1). We hope to deploy this into production

in 2020 after some final optimization of BacPipe processing. The pipeline is currently being prepared for publication by Xavier *et al.* Bacpipe: A Rapid, User-Friendly Whole Genome Sequencing Pipeline for Clinical Diagnostic Bacteriology and Outbreak ,Detection *Clinical Microbiology and Infection CMI* 2018.

Figure 4. BacPipe workflow for bacteriology clinical diagnostic and out outbreak detection.

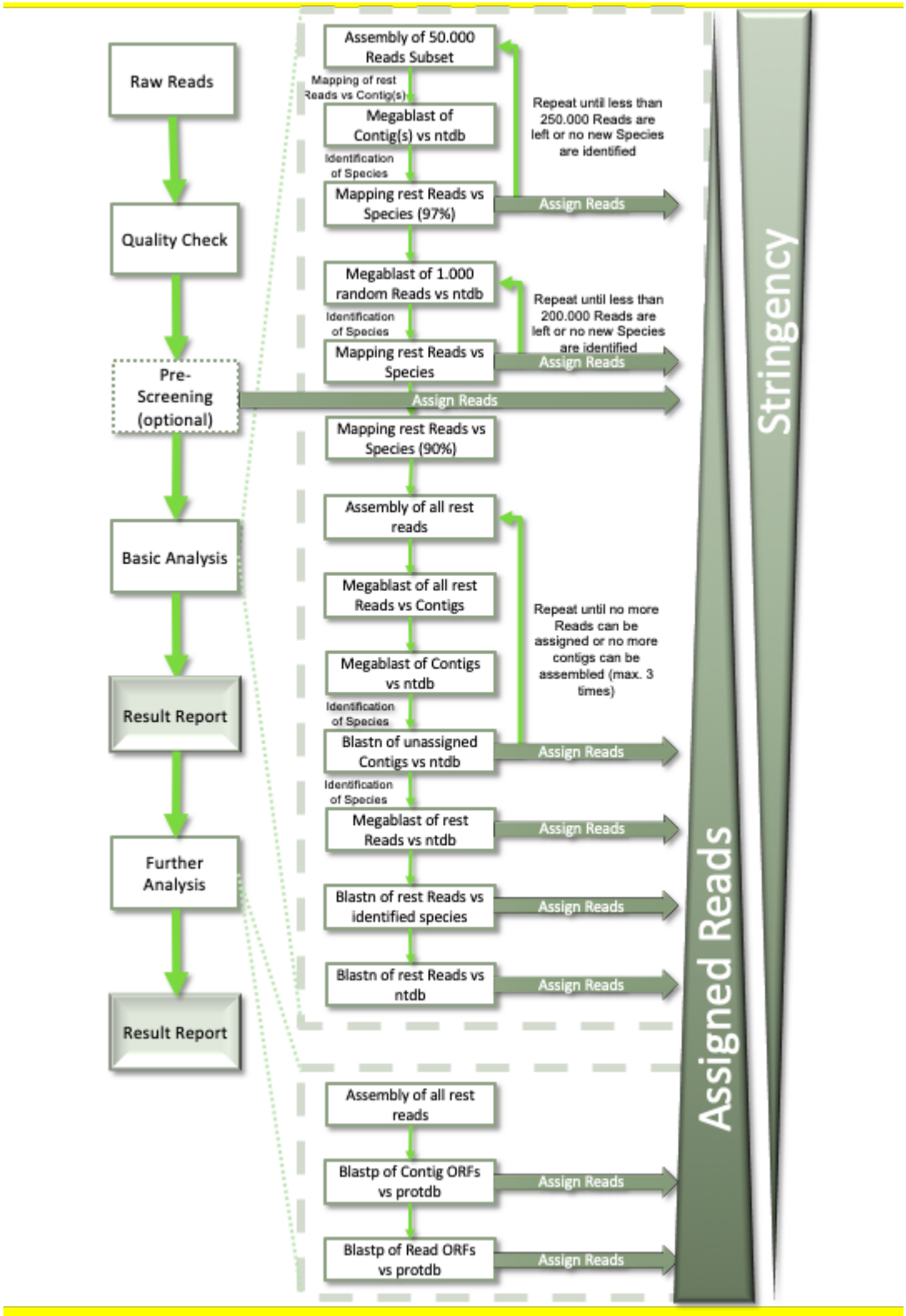


FLI RIEMS

The Reliable Information Extraction of Metagenomic Sequence data sets pipeline, developed by Friedrich Loeffler Institute (FLI-RIEMS) taxonomically classifies metagenomic sequence data with high accuracy. The pipeline uses a variety of analysis techniques, such as megablast and blastn, sequentially with decreasing stringency in assignment in order to classify all reads within a data set. The pipeline includes a compulsory basic analysis step and an optional further analysis stage, shown in Figure 5. The RIEMS workflow takes in genomic read files as input. Output results include a PDF of the results for taxonomic classification with R plots that present the results graphically. Additionally, tabular files provide a summary of the taxonomic classification results indicating the specific stage of the workflow the classification occurred as this reflects the specificity. The latest pipeline is currently available through a public GitHub repository (https://github.com/EBI-COMMUNITY/fli-RIEMS). EMBL-EBI is currently working with the developers of the pipeline in order to integrate with SELECTA. This involves the containerisation of the pipeline and optimisation for processing time and resource usage. The pipeline was recently published by Scheuch *et al.* RIEMS: a software pipeline for sensitive and

comprehensive taxonomic classification of reads from metagenomic datasets (2015) BMC Bioinformatics 16(69).

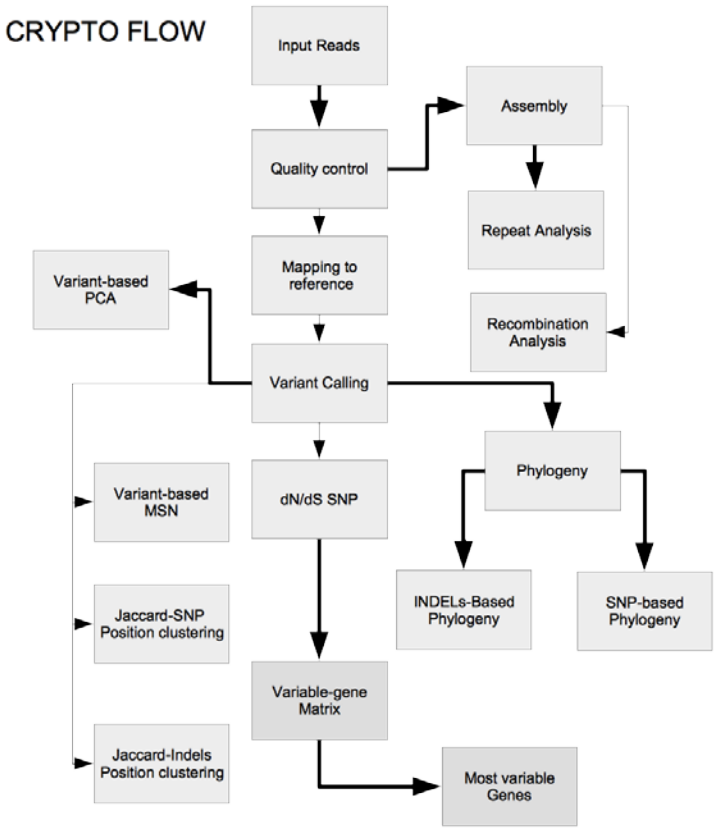
Figure 5. The RIEMS workflow, including both basic and further analysis stages.



Crypto Parasite pipeline

The Crypto Parasite workflow leverages a set of open source bioinformatic tools for comparative analysis of pathogen strains structural variations. Its analysis results provide insights into the pathogen evasion of the host immune response to infection. Ultimately the analysis can provide potential targets for drug development. To characterise and formalize key reports of the pipeline, we are conducting a pilot study of intra-genome and inter-genome structural variant comparison of the *Cryptosporidium hominis* strains IbA10G2. *C. hominis* is a waterborne parasite that causes Cryptosporidiosis commonly referred to as “Crypto”. Cryptosporidiosis is a diarrheal disease caused by the microscopic parasite Crypto, which lives in the intestine of humans and animals and is transmitted through an infected person or animal's stool. Crypto strains from an outbreak in the United Kingdom, Spain, Sweden and United States were used in this pilot study. The parasite pipeline is not yet integrated into the SELECTA environment as its repeat and recombination analysis components are still under active development. An abstract of the parasite pipeline pilot study was submitted for the 7th international Giardia and Cryptosporidium conference to be held at Rouen University, Normandy, France from the 23th to the 26th of June 2019. The source code of the Crypto Parasite pipeline can be accessed via GitHub (https://github.com/EBI-COMMUNITY/ebi-parasite).

Figure 7. Crypto Parasite workflow for comparative analysis of pathogen strains structural variations.



Integration into SELECTA workflow environment

Summary of the status of integration into the SELECTA workflow environment and developments over the last year:

DTU CGE: *Fully integrated*
The major new release of the DTU CGE pipeline (Version 1.1) is fully operational within the SELECTA framework. This now includes Core Genome Multilocus Sequence typing (cgMLST) support. This required all corresponding datahub reads to be re-analysed and former DTU CGE analyses to be suppressed as a consequence. This process was effectively and efficiently managed by the SELECTA workflow environment.

EMC SLIM: *Fully integrated*
Version 1.0 is fully operational within the SELECTA framework, a major bug-fix release required the re-analysis of all corresponding datahub reads and old analysis suppressed subsequently. Adaptations were required following a failed analysis because it exceeded the default 50Gb memory allocation.

UAntwerp BacPipe: *Fully integrated*
Version 1.0 is implemented in the SELECTA environment and as this pipeline, like DTU CGE, is designed for bacterial genome analysis, adaptations are required in the co-presentation of results as now two different analysis result sets will be available per run. EMBL-EBI collaborated intensively for the integration of BacPipe into SELECTA in order to port the native BacPipe graphical user interface tools into a command line version consumable by the automated SELECTA environment.

FFLI RIEMS: *Under integration development*
Currently benchmarking the FLI-RIEMS pipeline in preparation for integration into SELECTA. It currently takes an average 10 hours to process <1.5gb Fastq file, and we are collaborating with the developers to optimise time and resource usage, and it is still in need of further optimisation. Significant progress was made following the recent on site visit of the developer to EMBL-EBI.

Crypto Parasite Pipeline: *Under integration development*
The EMBL-EBI parasite pipeline is fully functional as a standalone pipeline, and is currently being prepared for integration into SELECTA once the current pilot project with Simone Caccio, the main stakeholder, is complete. The EMBL-EBI-Parasite pipeline was ported to the EMBL-EBI high performance computing (YODA) cluster for the pilot with all dependencies on open source bioinformatics tools installed therein. A pilot analysis of the *Cryptosporidium hominis* strains' genomes is ongoing.

Output from analytical workflows

Currently, three workflows have been integrated into the SELECTA workflow engine: DTU CGE, EMC SLIM and UAntwerpBacPipe. Figure 8 provides a snapshot of the number of submissions/analyses that were processed for each of these pipelines through SELECTA, showing the datahub from which analyses were generated. Figure 9 shows that within the last year there has been a constant increase in the number of runs submitted and processed through SELECTA with pipelines autonomously triggered regularly each month by SELECTA as new data are integrated into the data hubs. Note, the initial rise caused by the release of updated pipeline versions, which results in re-processing of all existing data. The versioning of existing pipelines and reprocessing of data is an important consideration for pipeline efficiency and automation.

Figure 8. Number of submissions/analyses processed per pipeline through SELECTA.

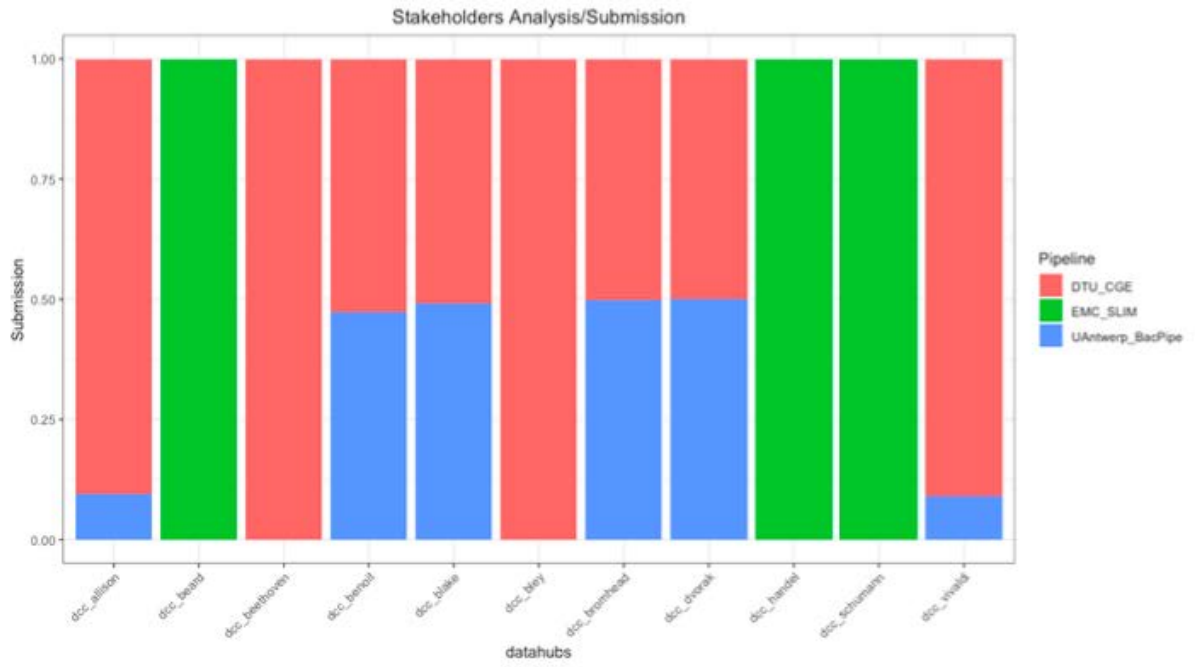
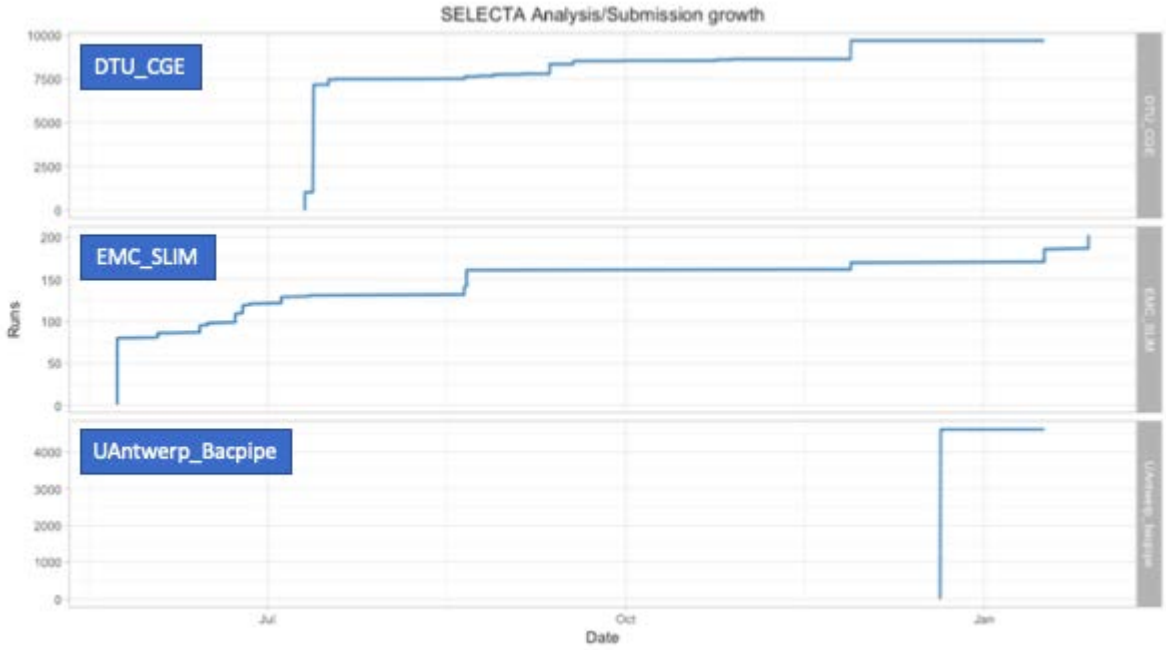


Figure 9. Number of runs process per pipeline in SELECTA over the last six months.



Future plans

- Integration and implementation of FLI RIEMS pipeline into SELECTA environment, requires containerisation of the pipeline and alterations to SELECTA to embed the pipeline in the environment
- Integration of Crypto Parasite pipeline into the SELECTA framework, currently involved in an ongoing pilot project in collaboration with Simone Caccio (Director, European Reference Laboratory for Parasites, Foodborne and Neglected Parasite Unit, Rome Italy), preparing the pipeline for integration in 2020.
- Possible adaptation of SELECTA environment to allow multi-stage analysis by consuming existing analysis objects in addition to raw reads from data hubs.
- Development of a SELECTA authenticated Application Programming Interface (RESTful API) to ease future remote management of the SELECTA workflow environment.
- Preparation of a containerised and deployable version of SELECTA workflow environment, so that it can be deployed and be scalable to cloud environments outside of EMBL-EBI.
- Preparation of a publication relating to SELECTA.

Update since reporting

- Final development and adaptation of the Crypto parasite workflow has been completed, including packaging of result outputs into user-friendly data structures, with the workflow ready for integration into SELECTA
- The Jovian workflow has been developed to address the identification and typing of viruses from mixtures, including important QC steps. Most recently, Jovian has been adapted for, and fully installed into, SELECTA.